BBC RD 1973/34

BBC • IBA

Engineering Research Report

The work covered by this report was undertaken by the BBC Research Department for the BBC and the IBA

The formation of a United Kingdom population data bank

R.E. Davies, M.A.

Research Department, Engineering Division
THE BRITISH BROADCASTING CORPORATION November 1973

BBC RD 1973/34
UDC 312.9
:681.3

THE FORMATION OF A UNITED KINGDOM POPULATION DATA BANK

R.E. Davies, M.A.

Summary

Data from the 1966 10% census for England, Wales and Scotland and the 1961 census for Northern Ireland were combined with data describing the density of inhabited buildings extracted from Ordance Survey maps to form a data bank of population figures within each half-kilometre grid square in the U.K. The required processing was carried out using the UCC bureau Univac 1108 computer and resulted in a computer-stored bank accessible by means of a programme to count populations within arbitrary contours.

Issued under the authority of

P. Rainger

Head of Research Department

Research Department, Engineering Division,
BRITISH BROADCASTING CORPORATION

November 1973
(RA-122)

THE FORMATION OF A UNITED KINGDOM POPULATION DATA BANK

THE FORMATION OF A UNITED KINGDOM POPULATION DATA BANK
R.E. Davies, M.A.

1. Introduction

The requirement for population counting in the UK arises in connection with the assessment of population coverage for existing and proposed television and radio services. The planning of these services results in the preparation of service area maps on which contours are drawn showing areas within which a satisfactory television or radio service can be received. It is important to determine the population within these areas and the net population within a chain of overlapping service areas carrying the same programmes.

Historically, service area population counting has been carried out manually for a number of years. Early efforts used the quoted populations of towns and counties which were proportioned and accumulated manually. In order to improve accuracy maps were prepared from the 1966 10% census showing the enumeration districts* and populations within them, the information having been bought from the census offices. By preparing transparent overlays of these it was possible to count populations within service areas proportioning partly-contained enumeration districts on an area basis. These results were thought capable of improvement in accuracy and, perhaps even more important, in repeatability; the manual methods were unable to repeat the same answers for any given area when counted by the different broadcasting organisations involved.

Only recently has it become possible to determine populations with any precision and repeatibility mainly because the enormous amount of data required for the accumulation of the population within large areas necessitates computer methods for its storage and access.

As the following example indicated, very high accuracy is not capable of achievement.

The area served by the Crystal Palace UHF television transmitters is approximately 10,000 sq kms and contains probably a population of 10 million people. An accurate count of population would involve a suivey of national census proportions and would accordingly occupy the attention of an army of surveyors for a considerable time. Clearly, such a survey would not be justified. Such data are not available from the census authorities, and are not likely to be available in the foreseeable future. Even supposing they were available, the cost of computer storage and retrieval would be prohibitively large. Typical estimated costs using up-to-date available computer services would be of the following order: £100 per day storage on rapid access medium or £40 per access if on magnetic tape.

A service for frequent use based on the above figures would be precluded by cost.

Clearly, the requirement must be met with a compromise between accuracy and access cost. The most convenient form in which to hold the data would be based upon a grid so that a population figure would be held within each of a set of grid squares covering the whole of the UK. The choice of size for the grid square will affect both the cost of data storage and retrieval, and the accuracy of population counts. The smaller the square the greater is the cost but the greater also is the accuracy.

As regards the accuracy, it is clear that population counts within small areas will be less accurate than population counts within large areas for statistical reasons, as explained below. An area will wholly contain some grid squares and partly contain others. Thus, upper and lower limits may be placed on the population for any area which will be the total population within all partially and wholly contained squares, and the total population within only wholly contained grid squares. For small areas we may only be able to bracket the population between zero and some figure. For large areas the upper and lower limits are likely to be in a reasonably close ratio. In general an intelligent guess may be made regarding the accuracy, based upon the ratio of the area to be counted and the area of the grid square.

The size of grid chosen, namely a half kilometre,† was partly in alignment with data banks of geographical features, e.g. ground height, being prepared for other purposes and partly an intelligently guessed compromise between the labour of assembling the computer data and the accuracy with which it can be used.

The populations within half km squares was not available from the census office and is not yet available in 1973*. The data bank had to be based on the enumeration districts and their populations for either the 1961 or 1966 censuses. The latter was chosen because it was thought to be less out-of-date, even though it would inevitably ne less accurate since only a 10% population sample was counted. Most enumeration districts were much larger than the half-km base square and the question of how to proportion the population of an enumeration district between half-km squares had to be studied.

*Enumeration districts are those areas accounted by an enumeration officer of the census. They vary considerably in size according to the population density, from only a few hundred metres in extent to several tens of kilometres. For the 1966 census between 40 and 45 thousand enumeration districts were used to cover England, Wales and Scotland.

*One-km square data are expected to be available later in 1974 based in the 1971 census. Hundred metre square data will also be available covering some of the U.K.

†These features, in addition to the concept of the populations data bank, are due to Mr. R.S. Sandell.

The method chosen involved spreading the population in proportion to the estimated area of inhabited buildings†, which was extracted from one-inch ordnance survey maps. There are many reasons why this would not give absolutely accurate proportioning. For example, the building area density (i.e. the proportions of the area shown on the map as covered by buildings) will only be approximately proportional to the population density; moreover it could only be estimated approximately if the data extraction was not to occupy many years work. Nevertheless, it is reasonable to suppose that this method will give much more accurate answers than treating the population as distributed evenly over each enumeration district, moreover, the population over areas covering many enumeration districts will retain the accuracy of the original data, since the total count within an enumeration district will be unaltered.

Since over a million half-km squares are needed to cover the UK, the processing necessary to achieve the data bank was carried out entirely by computer, requiring a very powerful system with substaintal storage capacity. Research Department had only limited experience of up-to-date facilities at the time when the data had been assembled. However, a service area prediction program was currently running at the University Computing Company (UCC) bureau in London, and it was clear that this bureau had sufficient available computing power and random-access storage capacity to carry out the initial processing and to house the data bank and access program when it was ready.

The data were assembled on paper tape (totalling about 40 kms in length) by the beginning of 1972 after some two years work. It was intended to complete the processing early in 1972; however, the problems caused by various kinds of error in the data were not fully appreciated and the first program was abandoned after several abortive attempts to run it. The problem was re-assessed and it was decided to process the data in much smaller protions at a time, with considerably more error checks. The new programs were written and run during late 1972 and early 1073 by which time much more experience had been gained of the UCC operating systems. The data bank was finally set up on their random access storage mediat in April 1973.

This report describes the stages necessary to form the data bank. Another report[1] describes the principles of the access program.

2. Method used to extract grid-based populations

It is required to apportion the population within an enumeration district among the set of half-km squares wholly and partly contained within it, in proportion to the area of inhabited buildings within each half-km square.

Suppose that the set of half-km squares are numbered from 1 to N so that S_i is the i^{th} square. Let B_i be the area of inhabited buildings within S_i and let A_i be the area of S_i which lies within the enumeration district. One must further assume that the inhabited buildings within a half-km square which is only partially contained within the enumeration district are evenly spread throughout that square so that the product A_iB_i is a measure of the inhabited buildings for S_i which are within the enumeration district. The proportioning factor for the distribution of population is therefore:

$$A_iB_i \,/ \sum_{j=1}^{N} A_jB_j$$

The total of all such factors is:

$$\sum_{i=1}^{N} A_iB_i \,/ \sum_{j=1}^{N} A_jB_j$$

which is clearly unity.

Hence the population P_i within square S_i may be taken to be:

$$P_i = E\,A_iB_i \,/ \sum_{j=1}^{N} A_jB_j$$

where E is the total population for the enumeration district. The right-hand side is in general a non-integral number but the left-hand side has to be integral. Consequently, this formula is rounded upwards or downwards to the nearest integer, a process which in practice was found to involve errors under 1%, in general, in the total populations for enumeration districts, since these were rarely less than a few hundred.

The individual populations within each half-km square and for each enumeration district are added together so that half-km squares which are partially contained within more than one enumeration district receives several populations.

Whilst other methods of apportioning population could have been arranged, this method has the advantages of being reasonably easy to understand and implement, and does ensure that the ultimate populations counted for very large areas will accord closely with the result which would have been attained by summing all the enumeration districts.

3. Preparation of the data for computer input*

The accumulation of all the required data describing enumeration districts, populations and inhabited building densities was a very substantial task which took about two years to complete and absorbed about 4 man-years of drawing office and operator effort. The entire data were assembled on eight-hole paper tape using manual keyboard punching for the inhabited buildings data and

*The procedure for preparing the data was due to Mr. R.S. Sandell, who also supervised the work.

automatic punching from a trace-digitiser for the enumeration district contours.

3.1. Inhabited buildings data

The data were extracted from one-inch to the mile ordance survey maps which show buildings shaded. A transparent overlay was prepared marking the 400 half-km squares within a 10 km grid square, and the operator was required to place the overlay in position over the map and estimate by eye the proportion of inhabited buildings to the nearest 10%. To assist with this process a key of squares was also drawn on the overlay showing exact proportions at 10% intervals up to 100%. The operator was asked to exclude buildings which one could class as uninhabited, such as churches, where one could clearly distinguish these from the map. Since, in practice, a large number of half-km squares contain a small but non-zero building area, a further category of square was also included described as 'nearly but not quite empty'. The building area density for this type of square was settled to be $3\frac{1}{3}$%.

Each estimated building density was recorded by the operator in a 20 x 20 table using the codes:

0	for zero
A	for the $3\frac{1}{3}$% category
1	for 10%
2	for 20%
3	etc.
* * *	
9	for 90%
T	for 100%

The operator also recorded on the sheet the 2-figure National Grid Reference of the bottom left-hand corner of the 10 km square, e.g. SH41.

Fig. 1 illustrates the process for a typical 10 km square namely NGR SW62. Fig.1(a) is the operator's view of the one-inch map (Ordance Survey Sheet 189) with the transparent overlay in position. The operator fills in the code for each ½km square on the overlay to produce the table, Fig. 1(b). Finally, the paper tape is punched giving the listing, Fig. 1(c).

3.2. Enumeration districts data

The enumeration district maps were obtained from the census office in microfilm form. These were first enlarged and retraced on a 1 inch = 1 mile scale. In densely populated cities (London, Glasgow etc.), where enumeration districts were small, a 2½ inch = 1 mile scale was used. The retraced maps were assembled in rolls, each roll covering a complete strip across the UK extending 30 km in a N–S direction with an overlap of 10 km between rolls. An example of a retraced map is shown in Fig. 2, which is printed at 2:1 reduction.

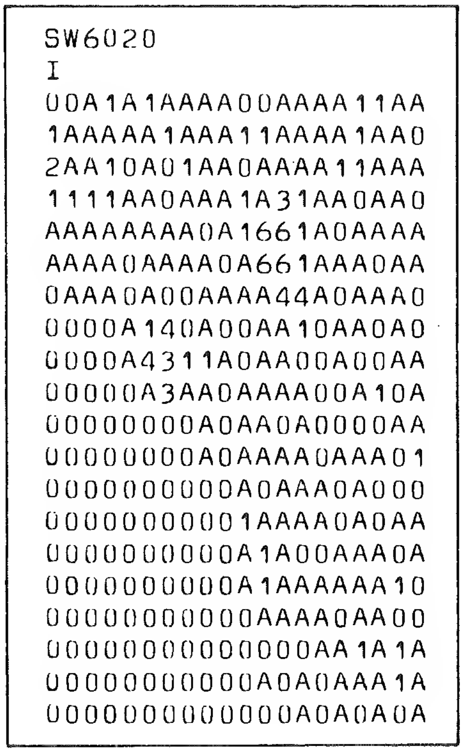
The enumeration districts were then traced over using a trace digitiser to produce the contour information on paper tape. The operator traced all enumeration districts within a 20 x 20 km area, preceding each by manually punched three-figure X and Y co-ordinates of the bottom-left corner of the 20 km square. Two reference points immediately followed the origin, and each enumeration district was terminated with a manually punched code which included the population for that district. It was also necessary for the operator to ensure that the trace digitiser scales were accurately set up and that 10 km on the map corresponded exactly with 400 units on the digitiser in both X- and Y- directions. Manual codes were also available for 'end of origin', 'cancel faulty tracing', and a few others.

The tracing proceeded from left to right (W to E) along each roll and from S to N taking each roll in turn.

In total, over 35,000 enumeration districts* were traced, producing about 36 km of paper tape. The total number of origins was about 800, covered in some 70 map rolls.

The first few kms of paper tape were checked for a few obvious types of error by a computer analysis and it was revealed that there was a drift in the trace digitiser which, it appeared, took some hours to settle after switching on. Some re-tracing was necessary, since this was not discovered until the tracing was some 20% completed.

*In London, the enumeration districts were often so small that several were merged into one larger district.



```
SW6020
I
00A1A1AAAA00AAAA11AA
1AAAAA1AAA11AAAA1AA0
2AA10A01AA0AAAA11AAA
1111AA0AAA1A31AA0AA0
AAAAAAAA0A1661A0AAAA
AAAA0AAAA0A661AAA0AA
0AAA0A00AAAA44A0AAA0
0000A140A00AA10AA0A0
0000A4311A0AA00A00AA
00000A3AA0AAAA00A10A
00000000A0AA0A0000AA
00000000A0AAAA0AAA01
0000000000A0AAA0A000
00000000001AAAA0A0AA
0000000000A1A00AAA0A
0000000000A1AAAAAA10
00000000000AAAA0AA00
00000000000000AA1A1A
00000000000A0A0AAA1A
0000000000000A0A0A0A
```

Fig. 1(c) - Listing of manually-punched paper tape for inhabited building density

Topographical information based upon the Ordnance Survey map with the sanction of the Controller of H.M. Stationery Office "Crown Copyright Reserved"

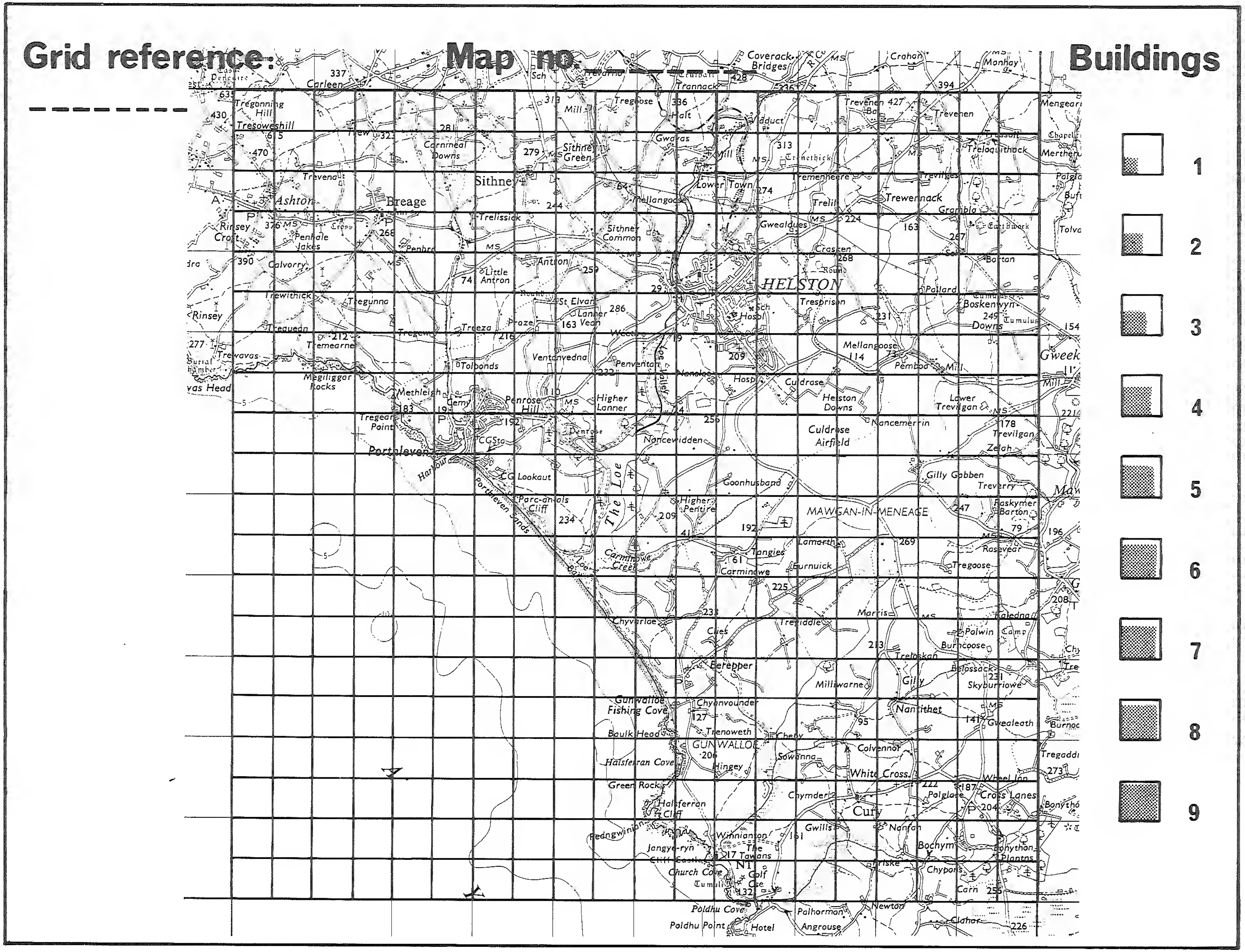


Fig. 1(a) - Operator's view of map with transparent overlay in position

Grid reference: SW 6020

Map no. 189

0	0	A	!	A	!	A	A	A	A	0	0	A	A	A	A	1	1	A	A
!	A	A	A	A	A	1	A	A	A	!	!	A	A	A	A	!	A	A	0
2	A	A	1	0	A	0	1	A	A	0	A	A	A	A	1	1	A	A	A
!	!	!	!	A	A	0	A	A	A	!	A	3	!	A	A	0	A	A	0
A	A	A	A	A	A	A	A	0	A	1	6	6	!	A	0	A	A	A	A
A	A	A	A	0	A	A	A	A	0	A	6	6	!	A	A	A	0	A	A
0	A	A	A	0	A	0	0	A	A	A	A	4	4	A	0	A	A	A	0
0	0	0	0	A	!	4	0	A	0	0	A	A	!	0	A	A	0	A	0
0	0	0	0	A	4	3	1	!	A	0	A	A	0	0	A	0	0	A	A
0	0	0	0	0	A	3	A	A	0	A	A	A	A	0	0	A	1	0	A
0	0	0	0	0	0	0	0	A	0	A	A	0	A	0	0	0	0	A	A
0	0	0	0	0	0	0	0	A	0	A	A	A	A	0	A	A	A	0	1
0	0	0	0	0	0	0	0	0	0	A	0	A	A	A	0	A	0	0	0
0	0	0	0	0	0	0	0	0	0	1	A	A	A	A	0	A	0	A	A
0	0	0	0	0	0	0	0	0	0	A	1	A	0	0	A	A	A	0	A
0	0	0	0	0	0	0	0	0	0	A	1	A	A	A	A	A	A	1	0
0	0	0	0	0	0	0	0	0	0	0	A	A	A	A	0	A	A	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	A	A	1	A	1	A
0	0	0	0	0	0	0	0	0	0	0	A	0	A	0	A	A	A	1	A
0	0	0	0	0	0	0	0	0	0	0	0	0	A	0	A	0	A	0	A

Buildings

▨	1
▨	2
▨	3
▨	4
▨	5
▨	6
▨	7
▨	8
▨	9

Fig. 1(b) - Coded building density table completed by operator

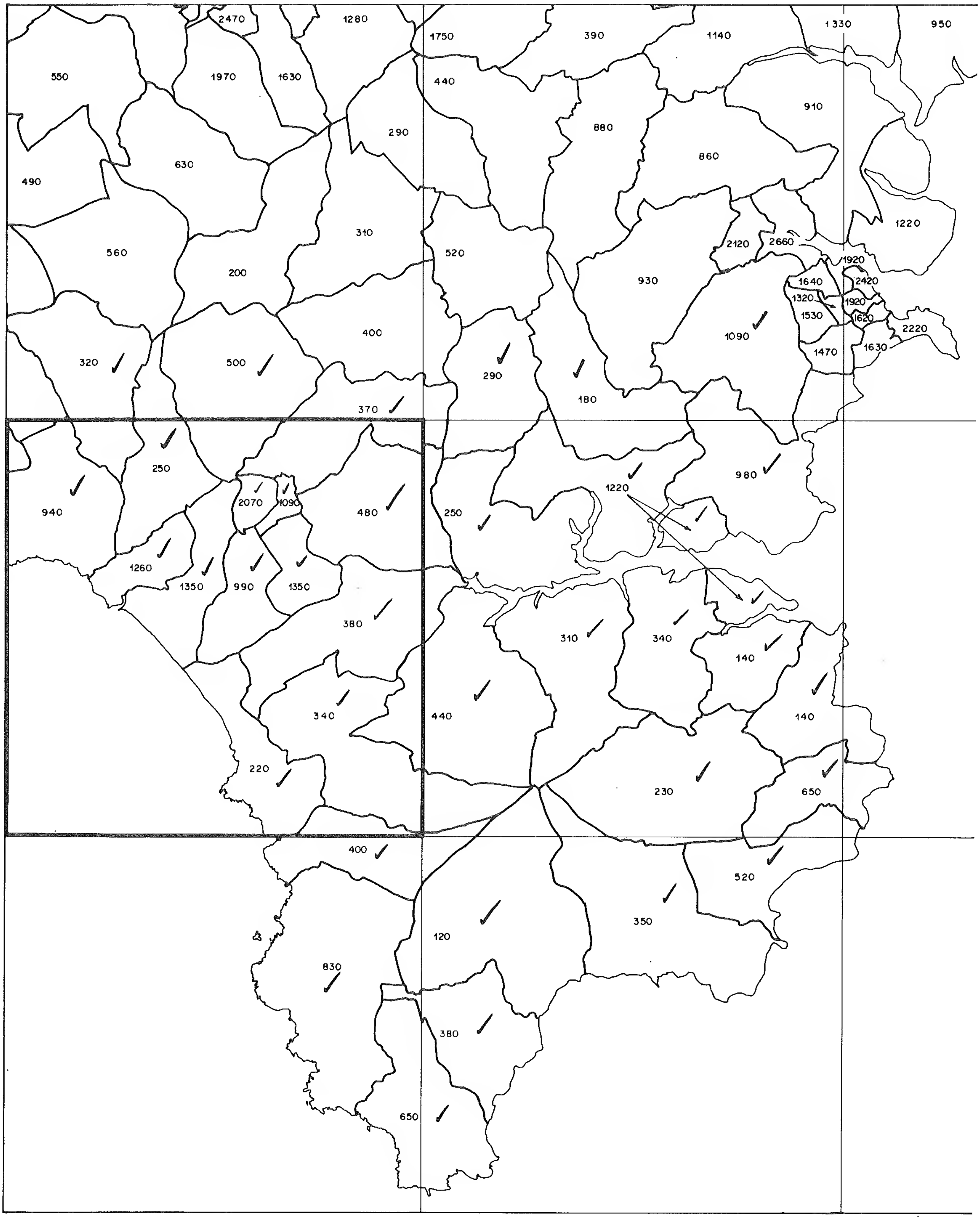


Fig. 2 - Section from enumeration districts map (reduced 2:1)
10 km square chosen for illustration in Fig. 1 is heavily outlined

4. Computer processing

It was originally intended for the sake of economy, to carry out the whole processing in a single long run totalling about six hours on the UCC 1108 computer. This run was, however, preceeded by a short sample run to ensure that there were no unforeseen difficulties. Unfortunately, the rate of detectable errors was higher than had been hoped, and other computer system problems compelled the abandonment of this approach. Difficulties also occurred because no terminal to the 1108 was locally available and it was necessary to have the paper tapes handled by UCC operators, who were unused to the large quantities involved. Since a Remote Job Entry computer with high-speed printer and paper tape reader was installed at about this time it was decided to carry out paper tape reading locally, the data being transmitted via private GPO line to the 1108 computer at the UCC centre.

The paper tape processing was planned in four stages. Firstly, the paper tape was read, checked and transmitted to magnetic tape files at the UCC centre. Secondly, these files were read with corrections merged, and each enumeration district was checked and processed to determine the list of half-km squares within it together with their populations, the inhabited buildings data file being read as required. Thirdly, this file was read to extract the populations and place them in correct order in the final population data file. Finally, enumeration districts remaining unprocessed because errors had been detected at the second stage were corrected, re-processed and entered into the file.

4.1. Paper tape reading and checking

This stage involved about 100 hours of computer terminal operation. It was found desirable to divide the paper tape into batches of approximately one km each. A total of 41 batches was read, checked and filed in separate runs. Each run produced a summary listing showing file position, origin, enumeration district population for the population data or NGR for the buildings data, together with a print-out of any lines with detected errors. The checks were for non-acceptable (e.g. non-numeric) characters and for incorrect number of characters per line.

The inhabited buildings data, consisting of five batches of paper tape, were corrected immediately by inserting fresh copies of the 10 km square records into the five files. These were then merged into a single file. Since the original order of punching had been in 100 km grid squares and the final desired order was to be in 10 km strips from West to East, this file had to be sorted. The final file consisted of 3050 10 km blocks of 20 x 20 characters occupying a total of about 270,000 computer words.

The enumeration district data consisted of 36 batches of paper tape and was read into 36 files of about 35,000 computer words each. The data were packed so that each word held both the X and Y co-ordinate values together with a key describing whether the word contained the co-ordinate of a contour point, the co-ordinate of an origin, or a population, or whether that line of data was in error.

4.2. Population processing

The extraction of grid-based populations in half-km squares from the enumeration district contours and populations and the inhabited buildings data was carried out by reading the contour for each enumeration district from the file already set up and first determining the maximum and minimum X- and Y-co-ordinates of a rectangle which just contained the enumeration district. A two-dimensional array was set up to receive the inhabited buildings; this contained the area and population for each half-km square of the rectangle. The contained area was evaluated by a sophisticated method involving only one pass through the contour points to determine the areas partially contained within the half-km square followed by one pass through the half-km square array to determine which of the remaining squares were wholly within the contour and which were wholly outside. It was essential to use a sophisticated technique because rough calculations had shown that a straightforward method would have considerably increased the running time, and therefore the cost, of the processing. To pick up from file the inhabited buildings data for each half-km square by random access read from a drum file would also have been far too expensive, and it was therefore necessary to allocate most of the computer core space to this data. This enabled direct access to data that would be required for a sequence of enumeration districts geographically separated by as much as 30 km in a N–S direction, since data for three ten-km strips were held in core together. Having set up in the array the area data and inhabited buildings data the determination of the proportioning factors for population distribution was straightforward. The whole information was written to magnetic tape files and a single line of outline information was printed for each enumeration district.

Before this stage could be completed for an enumeration district a number of checks were applied to the contour and population:

1) File sequence: the information from contour file was in the correct sequence and within specification where applicable.

2) Number of points: the number of points specified for a contour was greater than 2.

3) Contour closure: the first and last points were close together.

4) Contour size: the maximum extent in N–S and E–W directions did not exceed 30 km.

5) Area: the contour area was the same when computed by two independent methods (included mainly to check the program algorithms)

6) Population: the population was less than 10,000

7) Inhabited buildings: the total inhabited buildings within the contour was not zero

Checks (3) and (6) were relaxed in some areas, e.g. the populations for enumeration districts in London were sometimes greater than 10,000. If any of the above checks failed, that enumeration district could not be processed and the output listing contained error code numbers to indicate each failed district and the reason. Of the 35,000 enumeration districts almost 300 contained some kind of error. A special error file was set up to contain the outline information for these, so that subsequent runs could produce the corrected information. This process, however, was delayed until the final data bank was set up.

4.3. Final sorting runs

The process described above had produced 36 files of some 300,000 words each, stored on five magnetic tapes, and it was necessary to read these tapes and set up a final magnetic tape in which the geographically scattered population data for enumeration districts was sorted and merged into the geographically-ordered data for 10 km squares.

Three runs were carried out to merge the information for each enumeration district with the accumulating 10-km-squared-ordered data held temporarily on a random-access drum file of some 500,000 words. This file was read sequentially and stored on magnetic tape. In a further run, the three magnetic tape files were merged together to form a single file of 1·3 million words containing the whole data. At this stage one computer word was allocated for each half-km square; it contained the inhabited buildings data, the population, and the total area from all enumeration districts intersecting that square.

This file was now complete except for the 300 enumeration districts omitted earlier because errors had been detected. On account of mounting pressure to complete the job within the specified time and budget, no attempt was made to retrace these areas. Instead, it was assumed that the enumeration district consisted of all area which was as yet unprocessed within a rectangle containing the enumeration district. Thus, by reading the main file extracting the unprocessed area and inhabited buildings data, the population to be entered could be proportioned among half-km squares and added to the main file.

5. Discussion of population data errors

The title of this section is significant in that it has not been possible to analyse the errors. The essential problem is that there are no data against which to compare the results, and it would be as costly to produce such data as to set up the original. Thus, one's best hope of achieving accuracy was to carry out each stage as meticulously as possible, introducing error checks wherever it was possible and reasonable to do so.

Every effort was made to ensure that the manually-collected data were as accurate as possible but one cannot reasonably assume that the final data are 100% error free. The computer processing stages included every possible kind of error check. This was valuable not only to check the data, but also to assist with program development and to assure the programmer that the program did in fact work. Each error produced during the processing was subjected to close scrutiny and traced back to its source, if necessary through the paper tape back to the original maps. This was possible, because, as it turned out, the proportion of errors was reasonably low. It would hardly be possible to study several thousand such problems in detail in a reasonable time. A few computer runs were undertaken to attempt to detect errors but these were generally disappointing since, as suggested above, no really sound method of checking could be found. Of the methods tried, the most important was the area check, in other words it was intended to check the processed area for each half-km square which should, of course, be ·25 km^2. The main obstacle was the coastline, which intersects some 20,000 half-km squares, the sea areas of which had not been processed. The enumeration district tracing could have stretched well out into the sea to avoid the problem, but this requirement could hardly have been foreseen at the time when the tracing was done. Another check, however, was more encouraging. One could reasonably expect that enumeration districts with no population would contain no inhabited buildings and vice versa, and this proved to be the case for the few tens of enumeration districts to which the test could be applied.

In spite of the difficulty in checking it is still possible to be sure that counted populations in large areas will be closely in agreement with the census figures from which they were obtained because, although the single half-km square figure may be substantially in error the sums over enumeration districts have been preserved.

6. Form of data file for access purposes

Having extracted the population data in grid-based form it was necessary to produce a program to access the data and count populations within given contours. The details of this program are described in another report entitled: 'A population counting computer program' which is in course of preparation. The requirements of this program affect the form of storage of the population data which is discussed briefly in this Section.

The UCC computer bureau and most other computer complexes have two media available for the storage of large data files, both of which have disadvantages. Magnetic tape is a cheap medium but has to be read sequentially so that computer time is wasted in accessing data near the end of the file. Magnetic drum or disc is an expensive medium but random access to any part of the file is immediately available. Moreover, programs not requiring magnetic tape can be processed automatically by the computer operating system whereas magnetic tape programs require operators to load and unload the tapes. This affects the turn-round time of jobs so that non-tape jobs are processed much more quickly than tape jobs.

It was decided to hold the population data file on random access magnetic drum and reduce the storage cost as much as possible by compressing the data to minimum possible size.

A frequency distribution of the data was evaluated with the results tabulated below:

Population	No. of ½ km squares
0	783238
1 to 49	339322
50 to 10000	97440
	1220000

The simplest coding scheme possible would allocate one computer word to each ½ km square population. This would require 1·22 million words of storage. Observing that no population was larger than 10000 it would be possible instead to allocate 14 binary digits to each population thereby reducing the total number of words required to about 0·5 million. However, since nearly two-thirds of the populations are zero and 92% are less than 49 an alternative coding scheme was devised in which it was possible to code small numbers into a small number of binary digits at the expense of large numbers requiring more binary digits than would be necessary with a simple scheme. In outline the scheme is as follows: One binary digit is allocated per population to indicate whether or not it is zero. Up to three further groups of six bits determine the population.

Populations 1 — 48	require six bits
Populations 49 — 816	require twelve bits
Populations 817 — 16384	require eighteen bits.

With this scheme the total number of words used is approxiamtely 130,000. Since also the populations for ten km squares occupy varying space in the file a six-thousand word directory is also needed which relates ten km square to starting postion in the file.

7. Conclusion

A population data bank was set up in a form convenient for computer access to count populations within specified areas. The data were stored as a single file on random access magnetic drum in units of 10 km x 10 km. Each unit contains 400 populations on a half-km grid coded for minimum total file length. A directory indicates 10 km square number against starting position in the file. The data are ordered in 10 km West to East strips across England and Wales, the strips being ordered from South to North. Data for England and Wales are followed by data for Scotland and then the Channel Islands and finally Northern Ireland.

Access is by means of a program to count populations within specified contours; the access program is described in a separate report.

8. Reference

1. Research Department report in course of preparation 'Population counts using a computer data bank.'

CLA/AMM

Printed by BBC Research Department, Kingswood Warren, Tadworth, Surrey